A comparison of the alr and ilr transformations for kernel density estimation of compositional data

J.E. Chacón[1], **J.A. Martín-Fernández**[2] **and G. Mateu-Figueras**[2]
[1]Universidad de Extremadura, Cáceres, Spain
[2]Universitat de Girona, Girona, Spain

Abstract

In a seminal paper, Aitchison and Lauder (1985) introduced classical kernel density estimation techniques in the context of compositional data analysis. Indeed, they gave two options for the choice of the kernel to be used in the kernel estimator. One of these kernels is based on the use the alr transformation on the simplex S^D jointly with the normal distribution on $\mathbb{R}^{D-1}$. However, these authors themselves recognized that this method has some deficiencies. A method for overcoming these difficulties based on recent developments for compositional data analysis and multivariate kernel estimation theory, combining the ilr transformation with the use of the normal density with a full bandwidth matrix, was recently proposed in Martín-Fernández, Chacón and Mateu-Figueras (2006). Here we present an extensive simulation study that compares both methods in practice, thus exploring the finite-sample behaviour of both estimators.

Key words: kernel density estimation, compositional data, alr and ilr transformations.

1 Introduction

Kernel density estimation techniques are well-known nowadays. However, this method has been developed mainly for real univariate and multivariate data (see Wand and Jones, 1995), with only a few exceptions, as Hall, Watson and Cabrera (1987), where density estimation with spherical data is explored.

Another such exception is the seminal paper of Aithison and Lauder (1985), where kernel density estimation for compositional data is introduced. However, since this paper there have been many advances, both in techniques for manipulating compositional data, as the new ilr transformation proposed by Egozcue and others (2003) and also in kernel techniques, many of them related to the problem of the choice of the smoothing parameter, which is crucial for the good performance of the kernel estimator.

Many of these advances were recently compiled and expanded, from a theoretical point of view, in Martín-Fernández and others (2006). There, the use of Dirichlet kernels as well as alr and ilr Gaussian kernels fro density estimation is explored, together with the possibility of incorporating new bandwidth matrix selection procedures, as those in Duong and Hazelton (2003) or Duong and Hazelton (2005), to the kernel estimator.

In this paper we perform an extensive simulation study to compare how the different density estimation methods for compositional data perform in practice.

2 Kernel density estimation for compositional data

2.1 The estimators

Given compositional data $X_1, \ldots, X_n$, coming from an absolutely continuous distribution on the simplex S^D, with density $f : \mathcal{S}^D \to \mathbb{R}$, following Aitchison and Lauder (1985) we define the *kernel estimator* of f as

$$f_{nH}(x) = \frac{1}{n} \sum_{i=1}^{n} k(x|X_i, H), \quad x \in \mathcal{S}^D,$$

where the bandwidth matrix H is a positive definite matrix and the kernel $k(\cdot|X_i, H)\colon S^D \to \mathbb{R}$ is a density function on $\mathcal{S}^D$ centred on the data point X_i and spread out depending on the shape and size of the smoothing factor H.

Here we will focus on alr and ilr normal kernels. These are defined, respectively, as

$$k_{\mathrm{alr}}(x|X, H) = \phi(\mathrm{alr}(x)|\mathrm{alr}(X), H), \quad x \in S^D,$$

$$k_{\mathrm{ilr}}(x|X, H) = \phi(\mathrm{ilr}(x)|\mathrm{ilr}(X), H), \quad x \in S^D,$$

where $\phi(\cdot|\mu, \Sigma)$ is the density of the normal $N_{D-1}(\mu, \Sigma)$ distribution and $\mathrm{alr}(x)$ and $\mathrm{ilr}(x)$ stand for the additive log-ratio and isometric log-ratio transformations (see Egozcue and others, 2003), given by

$$\mathrm{alr}(x) = \big[\ln(x_1/x_D), \ldots, \ln(x_{D-1}/x_D)\big],$$

$$\mathrm{ilr}(x) = (y_1, \ldots, y_{D-1}) \in \mathbb{R}^{D-1}, \text{ with } y_i = \frac{1}{\sqrt{i(i+1)}} \ln \left(\frac{\prod_{j=1}^{i} x_j}{(x_{i+1})^i} \right).$$

The proposal of Aithison and Lauder (1985), labelled AL, consists of using the kernel estimator f_{nH} with the kernel k_{alr}. For the bandwidth matrix, they suggest to restrict its form to be $H = \lambda T$, where $\lambda > 0$ and T the sample covariance matrix of the additive log-ratio compositions

$Y_1 = \mathrm{alr}(X_1), \ldots, Y_n = \mathrm{alr}(X_n)$; that is, denoting $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$,

$$T = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})(Y_j - \bar{Y})^T.$$

Then, the choice of λ is made by maximizing the pseudo-likelihood function

$$\mathrm{PL}(\lambda) = \prod_{i=1}^{n}\Big\{\frac{1}{n-1}\sum_{j\neq i} k_{\mathrm{alr}}(X_i|X_j, \lambda T)\Big\}.$$

We should note that Aitchison and Lauder (1985) themselves highlight the problem that, using the alr transformation, it is only possible to work with a bandwidth matrix proportional to the sample covariance matrix of the alr-transformed data, due to the fact that the results may not be invariant under permutations of the components.

However, in real spaces, Wand and Jones (1995, p. 106) state that this parametrization of the bandwidth matrix is appropriate only for multivariate normal alr compositions but not for general density shapes. On the contrary, using the ilr normal kernel as proposed in Martín-Fernández and others (2006), all parameterizations of the bandwidth matrix H are feasible.

Therefore, we propose to use the kernel estimator f_{nH} with the ilr kernel, as it admits all possible parametrizations of the bandwidth matrix. In this sense, we will label this method with CV when the bandwidth matrix H is of full type (i.e., positive definite with no restrictions) and chosen via cross-validation; see Duong and Hazelton (2005). And we will label this method with DH if, following the recommendations in Duong and Hazelton (2003), we select a full bandwidth matrix using its SAMSE plug-in procedure.

2.2 Simulation setup and results

To compare the three density estimation methods we study their performance on estimating 12 test densities, whose ternary contour plots are depicted in Figure 1 below. These densities are closely related to the test densities appearing in Chacón (2008).

From each test density f we have generated 500 simulation samples of size $n = 100$ and, for each of these samples, we have computed the three density estimates, f_{CV}, f_{DH} and f_{AL} and their Integrated Squared Errors (ISEs), defined as $ISE(f_{\mathrm{CV}}) = \int(f_{\mathrm{CV}} - f)^2$, $ISE(f_{\mathrm{DH}}) = \int(f_{\mathrm{DH}} - f)^2$ and $ISE(f_{\mathrm{AL}}) = \int(f_{\mathrm{AL}} - f)^2$, respectively. The box-plots of the distributions of these ISEs for each method and each density in the simulation study is shown in Figure 2.

By looking at Figure 2 we immediately notice the well-known fact that the CV method is usually very variable. Precisely, in this situation it always has more variability than the other two methods. We should say, however, than in average terms the CV method has a very good performance, with a median ISE value that is sometimes below that of the DH and the AL methods, as it happens for density #12, for instance. That is, CV is a good method in average terms, but it is not quite trustworthy.

The AL method is far less variable than the CV method, although for some densities it provides completely wrong estimations; see, for instance, the results for densities #3 and #4.

Overall, our preferred method, in view of the simulation results, is DH, the ilr kernel method with a full bandwidth matrix chosen by a SAMSE plug-in procedure. It is often the least variable method out of the three of them, and it is never corrupted in the sense of always having a good average performance as well.

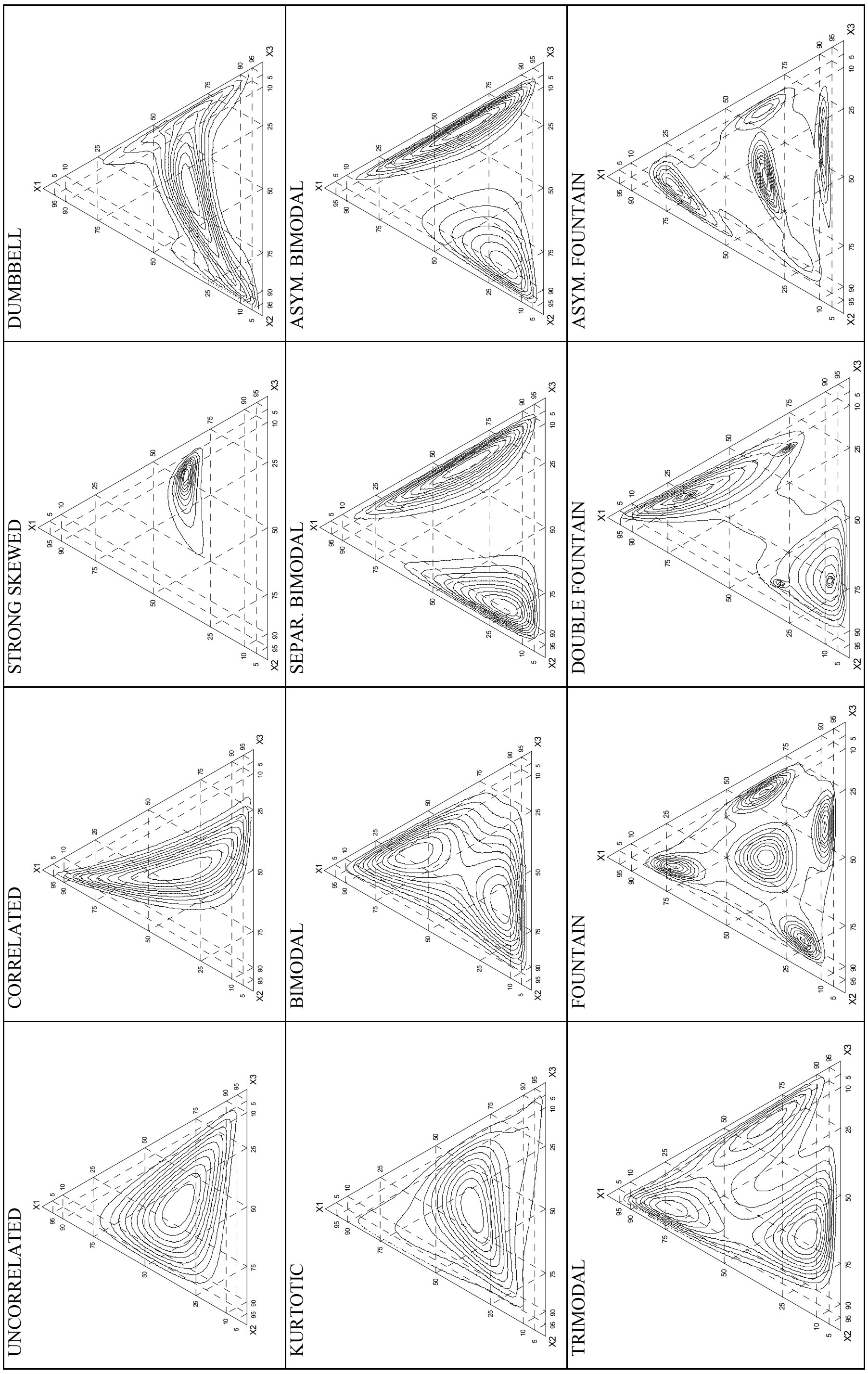

Figure 1: *Contour plots for the 12 test densities.*

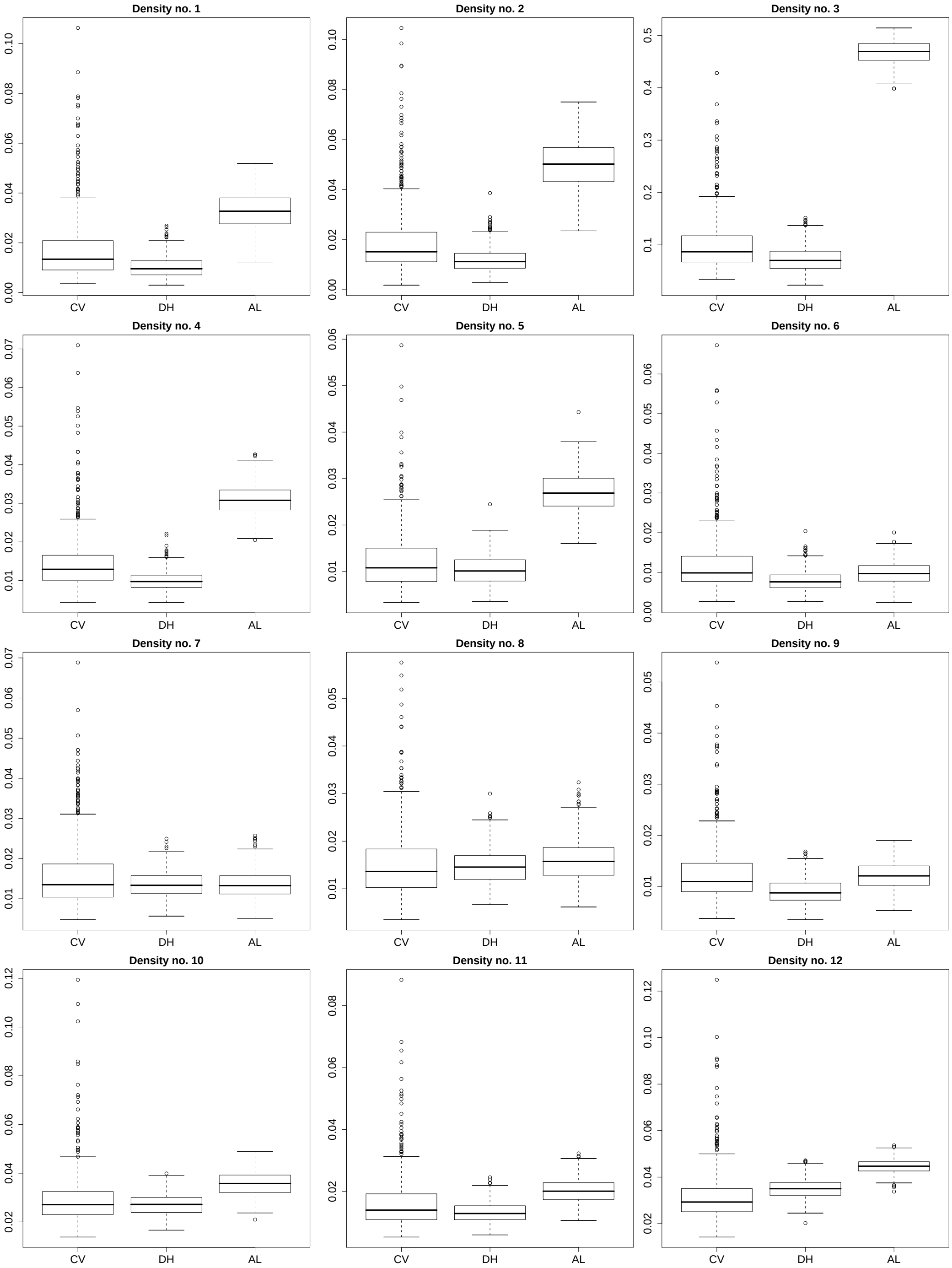

Figure 2: *ISE box-plots for the 12 test densities.*

Acknowledgements

This work has received financial support from the Dirección General de Investigación of the Spanish Ministry for Science and Technology through the projects MTM2006-06172 (first author) and BFM2006-03040 (second and third authors).

References

Aitchison. J. and Lauder, I.J. (1985). Kernel Density Estimation for Compositional Data. *Applied Statistics*, **34(2)**, 129–137.

Chacón, J.E. (2008). Data-driven choice of the smoothing parametrization for multivariate kernel density estimators. *Submitted.*

Duong, T. and Hazelton, M.L. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, **15**, 17–30.

Duong, T. and Hazelton, M.L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, **32**, 485–506.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003) Isometric log-ratio transformations for compositional data analysis. *Math. Geol.*, **35 (3)**, 279–300.

Hall, P., Watson, G.S. and Cabrera, J. (1987) Kernel density estimation with spherical data. *Biometrika*, **74**, 751–762.

Martín-Fernández, J.A., Chacón, J.E. and Mateu-Figueras, G. (2006). Updating on the kernel density estimation for compositional data. In A. Rizzi and M. Vichi (Eds.), *COMPSTAT 2006 Proceedings*, 713–720.

Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing.* Chapman & Hall, London.